

# SPEAKER ADAPTATION AND ENVIRONMENTAL COMPENSATION FOR THE 1996 BROADCAST NEWS TASK

*Vipul N. Parikh, Bhiksha Raj, and Richard M. Stern*

Department of Electrical and Computer Engineering & School of Computer Science  
Carnegie Mellon University  
Pittsburgh, Pennsylvania 15213, USA

## ABSTRACT

In this paper we present and discuss the performance of the Maximum Likelihood Linear Regression (MLLR) adaptation algorithm for various environmental conditions presented in the 1996 ARPA Hub 4 Broadcast News task. We also present a comparison of the effects of cepstral compensation using Codeword Dependent Cepstral Normalization (CDCN) and MLLR on the likelihoods of speech data and on the corresponding recognition error rates. Finally we describe a possible extension of MLLR using a quadratic regression equation.

## 1. INTRODUCTION

Broadcast news data is a highly varied domain with speech ranging from very high-quality recordings made in the broadcast studio to extremely noisy speech recorded on the street. Consequently, the recognition of broadcast news speech is a complex task requiring a speech recognition system that is robust to environment and speaker variations. A number of compensation techniques may be considered to handle the diversity of the task.

The 1996 ARPA Hub 4 task [1] consisted of speech recorded from a variety of radio and television shows, including conditions such as clean studio-recorded speech, spontaneous speech, speech with music in the background, speech over telephone channels, speech in a variety of noisy environments, and foreign-accented speech. In the case of the partitioned evaluation (PE), the task was made somewhat easier by manually pre-partitioning the data according to the conditions present in the speech; in the case of the unpartitioned evaluation (UE), this information was not available.

We compared the performance of the Maximum Likelihood Linear Regression (MLLR) adaptation algorithm [2] and the Codeword Dependent Cepstral Normalization (CDCN) compensation algorithm [3] as a means of improving the recognition performance on the various acoustic conditions present in the 1996 Hub 4 data. In this paper we report the results of these experiments. In addition, we also briefly describe the results of experiments with a Maximum Likelihood Non-linear Adaptation (MLNA) algorithm for adaptation of the HMM means.

In Section 2 we report and discuss the performance of the MLLR adaptation algorithm for various environmental conditions presented in the 1996 ARPA Hub 4 task. In Section 3 we describe the results of our experiments using a combination of CDCN compensation and MLLR adaptation. In Section 4 we describe the implementation of the MLNA algorithm which incorporates a nonlinear (quadratic) warping function, and we describe its performance. Finally, in Section 5 we present our conclusions.

## 2. PERFORMANCE OF MLLR ADAPTATION

MLLR [2] is an algorithm that adapts the means of the HMM distributions to model the data to be recognized more accurately. The adaptation could be either supervised (where the transcripts of the adaptation data are available), or unsupervised (where the adaptation transcripts are automatically generated). In either case it is assumed that the adaptation data and the speech being recognized are acoustically similar so that the adapted mean is truly representative of the data to be recognized. It is therefore useful, and perhaps necessary, for all the data that is being used for adaptation and the data that is being recognized to have similar characteristics (*i.e.* belong to the same environmental condition and preferably to the same speaker or group of speakers).

In the case of broadcast news, the speakers, the recording environment, and the quality of the speech are varying constantly. As a result it becomes necessary to segment the speech into regions of uniform condition. However, a single segment of data may be insufficient for purposes of adaptation. Consequently, the segmented data must be grouped together to provide clusters of sufficient duration for adaptation. The best possible clustering mechanism would be to use the condition and speaker labels provided with the data to group similar segments together. A good automatic clustering mechanism would result in performance similar to that obtained from clusters formed by using the labels provided.

Both segmentation and clustering for our experiments were performed by a relative cross-entropy based system [4]. For the PE, task segments and clusters were formed on data belonging to the same condition as given by the labels provided for that evaluation with the speech data. In the case of the UE, the segmentation and clustering were done over all the data.

We establish the importance of clustering for MLLR adaptation by comparing the result of adaptation without clustering and the result of using ideal clusters as given by the labels provided. We also consider the effect of the clustering approach used by our system in the context of these two results. Additionally, we evaluate the effect of adaptation using the automatically-generated clusters on the recognition accuracy obtained for the various types of speech present in the 1996 Hub 4 task. The baseline system consisted of a fully-continuous HMM trained on the Wall Street Journal (WSJ) SI-321 data. All experiments were performed on the development test (devtest) data provided by NIST. Results for the evaluation test data are included here for comparison purposes.

In order to evaluate the performance of various clustering approaches we performed a series of experiments on a subset of the sentences from the developmental test set representing the F0 condition (high-quality studio speech). Our baseline recognition

system using the unadapted means resulted in a word error rate (WER) of 21.3% for this subset of the F0 data in the devtest. In Table 2 we show the improvement obtained over the baseline WER by adapting to individual segments of speech, ideal clusters of speech (using information provided by the speaker labels and segment labels from the PE), and unsupervised automatically-generated clusters obtained using the clustering algorithm described in [4]. In addition, we compare recognition error rates obtained using “ideal” information directly from the actual transcript of speech with the imperfect hypothesis provided by the decoder itself.

The results summarized in Table 2 clearly demonstrate that the use of clustering improves speech recognition error rate. It is also apparent that the unsupervised automatic clustering algorithm is quite effective. In fact, the use of automatic clustering reduces the error rate compared to the error rate obtained with no clustering by two-thirds the reduction obtained with ideal clustering.

Transcript Source	Clustering Scheme	WER	Relative Gain
Ideal	None	12.2%	42.7%
Ideal	Ideal	17.6%	17.3%
Decoder	None	19.8%	7.0%
Decoder	Unsupervised	18.2%	14.5%
Decoder	Ideal	17.4%	18.3%

**Table 1.** Comparison of word error rates (WER) obtained for a subset of the development test set using MLLR adaptation with different clustering approaches and with different transcript sources. Word error rate for this dataset with no adaptation was 21.3%

Table 1 also includes results of control experiments where the adaptation of the HMMs was done on the testing data using the actual transcriptions. Not unexpectedly, the resulting error rate obtained by adapting on individual segments is extremely high. However, it is surprising to note that the availability of perfect transcriptions for clustered data results in no gain over having automatically- obtained transcriptions. Hence it appears that further improvements to clustering performance (rather than decoder performance) are needed to obtain further improvement in the effectiveness of the MLLR adaptation.

Table 2 shows the relative gain obtained from MLLR adaptation for other PE conditions in our experiments using subsets of the development test data. The clusters used for these results were all automatically generated. We observe that except for the cases of spontaneous speech (F1) and telephone speech (F2) the relative gain obtained by MLLR adaptation of the means is similar for all conditions. The reduced gain from MLLR adaptation observed in the case of telephone speech may be a consequence of the reduced bandwidth analysis used to generate the cepstra and the models. It is not immediately apparent why spontaneous speech did not benefit as much from adaptation as read speech. One possible hypothesis is that the difference in pronunciation between spontaneous

Condition	Baseline	Adapted	Relative Gain
F1	28.7%	27.3%	4.8%
F2	61.5%	57.5%	6.5%
F3	55.1%	47.1%	14.5%
F4	41.3%	35.3%	14.5%
F5	43.2%	36.8%	14.8%
FX	71.8%	62.8%	12.5%

**Table 2.** Comparison of WER obtained with and without MLLR adaptation for subsets of the development test data from six PE conditions.

speech and read speech is not adequately represented by the dictionaries being used for recognition, and that adaptation of means cannot account for such mismatches.

The FX data are speech segments where a combination of the various other conditions (such as foreign accented speech and noisy recording conditions) occur simultaneously. In the case of speech from the FX condition, we also ran experiments using the ideal clusters as given by the labels. It was observed that ideal clustering provided marginally worse WER than the automatic clustering. This appears to indicate that in the presence of multiple conditions of variation, the artifice of hand labelling of data holds no great advantage over the use of automatic methods to cluster acoustically-similar segments.

PE Condition	Baseline WER	WER after Adaptation	Relative Gain
F0	27.2%	26.1%	4.0%
F1	32.4%	32.3%	0.3%
F2	43.2%	39.7%	8.1%
F3	43.3%	37.3%	13.8%
F4	45.7%	43.9%	3.9%
F5	45.8%	38.1%	16.8%
FX	61.8%	57.8%	6.4%

**Table 3.** Comparison of WER obtained with and without MLLR adaptation for the evaluation test set data under the conditions of the Partitioned Evaluation (PE).

Tables 3 and 4 show the relative improvements in WER obtained for the PE and UE evaluation tests using MLLR adaptation of means in the various conditions. The improvements obtained for the various conditions were in general less than the corresponding improvements obtained for the development test data. This may be

a consequence of the fact that the baseline WER was generally higher than that obtained for the development test data. From the results for spontaneous speech (F1) in Tables 2, 3 and 4, it appears, once again, that MLLR adaptation is not reliable for this condition. Except in case of speech with music in the background (F3), the final WERs in the PE and UE obtained after MLLR adaptation are similar. It appears that the poor performance obtained for speech with background music in the case of the UE was because a large number of these segments were erroneously clustered with other types of data.

UE Condition	Baseline WER	WER after Adaptation	Relative Gain
F0	26.0%	24.8%	4.6%
F1	33.5%	33.7%	-0.2%
F2	44.7%	39.8%	10.9%
F3	48.4%	48.8%	-0.4%
F4	45.0%	42.5%	5.5%
F5	40.8%	38.8%	4.9%
FX	62.9%	60.3%	4.1%

**Table 4.** Comparison of WER obtained with and without MLLR adaptation for the evaluation test set data under the conditions of the Unpartitioned Evaluation (UE).

### 3. COMPARISON OF MLLR WITH CDCN

Traditionally CMU has applied the CDCN algorithm [3] to improve the robustness of speech recognition systems with respect to variable recording conditions. CDCN is a Maximum Likelihood-based algorithm that compensates cepstra in the testing environment for the effects of unknown additive noise and unknown linear filtering. Since the compensation is performed on incoming cepstra, the use of CDCN does not preclude an additional adaptation step such as using MLLR. In other words, it is possible to use MLLR to adapt the means to the CDCN-compensated cepstra instead of to the regular uncompensated cepstra.

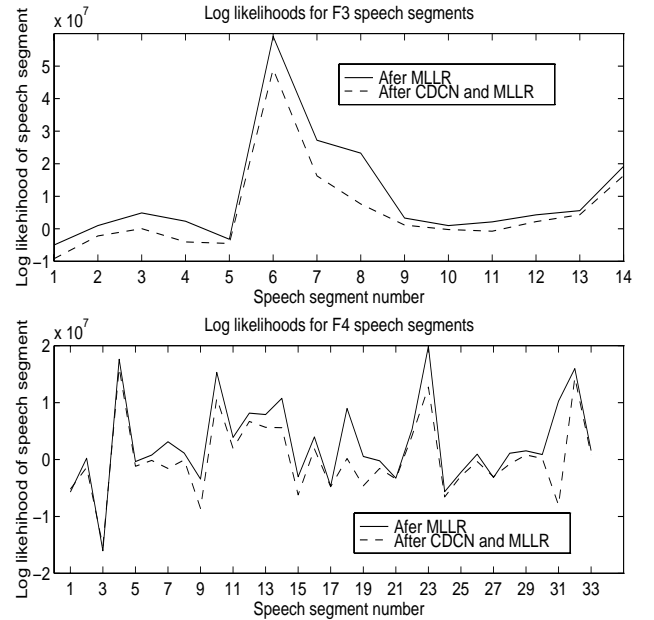
Condition	Baseline	After CDCN	After MLLR	CDCN+ MLLR
F3	55.1%	52.4%	47.1	47.3%
F4	41.3%	39.9%	35.3%	36.6%

**Table 5.** Comparison of WER obtained for subsets of the development test set using CDCN with and without MLLR.

Table 5 presents the results of CDCN compensation, MLLR adaptation and the combination of the two on speech with background music and with background noise. While CDCN by itself is effective, it is not as effective as MLLR. Furthermore, the addition of

MLLR to a system that already incorporates CDCN does not provide a lower WER than a system that uses MLLR alone.

Since both CDCN and MLLR are Maximum Likelihood-based algorithms it is instructive to look at the likelihoods of the observations after MLLR adaptation, and after compensation using the combination of MLLR and CDCN. We observe from Figure 1 that the likelihoods achieved by the combination of CDCN and MLLR are in fact *lower* than those achieved by MLLR alone.



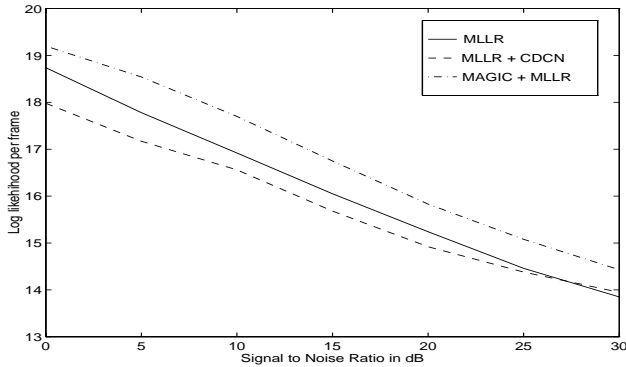
**Figure 1.** Comparison of log likelihoods of several segments of F3 and F4 speech after MLLR adaptation of means and after CDCN followed by MLLR adaptation.

To confirm this observation additional simulations were run on a number of speech files that were artificially corrupted with noise. The original “clean” speech was represented by a Gaussian mixture distribution. MLLR was used to adapt the means of this distribution. An alternate implementation of CDCN, referred to in this paper as “MAGIC”, adapt the means of the distributions of the HMMs rather than the incoming feature vectors. Figure 2 plots the log-likelihood of the data for a typical utterance after the models have been adapted using MLLR, CDCN followed by MLLR, and MAGIC.

We observe from these simulations that at lower SNRs the final likelihood achieved by MLLR alone is typically higher than the likelihood obtained by the combination of CDCN and MLLR. This agrees with our observations on the broadcast data. The best likelihoods obtained, however, are from the combination of MAGIC and MLLR. While the reason for this behaviour is not clear, this provides reason to believe that the use of MAGIC for the broadcast news data (either alone or in combination with MLLR) would provide further reductions in WER.

### 4. NONLINEAR ADAPTATION

MLLR adaptation uses linear regression to obtain speaker-specific and environment-specific means from speaker independent means.



**Figure 2.** Log likelihood of speech corrupted by white noise as a function of SNR after adaptation with MLLR, CDCN followed by MLLR, and MAGIC.

One interpretation of the relation between the adapted means and the unadapted means could be that MLLR provides a truncated version of a much longer series that actually represents the transformation from speaker independent means to the speaker/environment-specific means. One could therefore add additional terms to this series with the appropriate coefficients to obtain a more accurate adaptation. We extended the linear regression used by MLLR to include quadratic terms. In other words, the mean vectors of the Gaussian distributions were adapted using a quadratic transform as follows

$$\mu_k^{new} = A + B\mu_k + C\mu_k^2$$

where  $B$  and  $C$  are matrices of size  $n \times n$ ,  $A$  is a vector of size  $n \times 1$ , and  $\mu_k$  is the  $k^{th}$  mean of a mixture component (*i.e.* a component of the initial mean vector) of size  $n \times 1$ . The values of  $A$ ,  $B$ , and  $C$  are chosen to maximize the likelihood of the adapted models generating the adaptation data.

The parameters  $A$ ,  $B$ , and  $C$  are estimated in a manner very similar to that presented in [2] to estimate the regression terms for MLLR.

To evaluate the performance of the nonlinear algorithm, referred to as MLNA, an initial experiment was done using the speaker-dependent portion of the ARPA Resource Management (RM) database. Using training and testing data from 12 different speakers, a set of 40 sentences was used for adaptation and the remaining 60 were used for testing. The average result is reported in Table 6.

Baseline	MLLR	MLNA
8.0%	5.9%	5.4%

**Table 6.** Comparison of WER obtained using MLLR and quadratic nonlinear adaptation (MLNA).

We observe that while the use of the quadratic term appears to result in a consistent improvement over using the linear term alone, the improvement obtained is relatively small. The linear term seems to be capturing most of the information in the transformation, so that the improvement obtained from any subsequent terms appears to be small. Experiments performed on broadcast news data with MLNA adaptation showed similar results.

## 5. SUMMARY AND CONCLUSIONS

While the MLLR algorithm is effective for both environment and speaker adaptation, it works best when there is of acoustically-similar data available for adaptation. We found that the clustering algorithm described in [4] can provide a significant improvement in recognition accuracy.

We found that the addition of MLLR to a system that already incorporates CDCN on the ARPA Hub 4 task does not provide a lower WER than a system that uses MLLR alone. Nevertheless, preliminary findings indicate that the implementation of MAGIC, an algorithm similar to CDCN that models the means of the internal models of an HMM, is likely to be more promising.

Extrapolating the MLLR regression to include quadratic terms appears to result in no additional improvement for the original Resource Management database. However, this result remains to be confirmed using datasets with a wider variety of speech conditions.

## ACKNOWLEDGEMENTS

This research was sponsored by the Department of the Navy, Naval Research Laboratory under Grant No. N00014-93-1-2005. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government.

## REFERENCES

1. R. M. Stern. "Specification of the 1996 Hub 4 Task", elsewhere in these Proceedings.
2. C. J. Leggetter, and P. C. Woodland, "Speaker Adaptation of HMMs using Linear Regression", CUED/F-INFENG/TR. 181, June 1994, Cambridge University Engineering Department.
3. A. Acero. "Acoustical and Environmental Robustness in Automatic Speech Recognition", Ph.D. thesis, Department of Elect. and Comp. Eng., Carnegie Mellon University, 1990.
4. M. Siegler, U. Jain, B. Raj, and R. M. Stern, "Automatic Segmentation, Classification, and Clustering of Broadcast News Audio", elsewhere in these Proceedings.